Detecting structural variations in the human genome using next generation sequencing

Ruibin Xi, Tae-Min Kim and Peter J. Park

Advance Access publication date 6 January 2011

Abstract

Structural variations are widespread in the human genome and can serve as genetic markers in clinical and evolutionary studies. With the advances in the next-generation sequencing technology, recent methods allow for identification of structural variations with unprecedented resolution and accuracy. They also provide opportunities to discover variants that could not be detected on conventional microarray-based platforms, such as dosage-invariant chromosomal translocations and inversions. In this review, we will describe some of the sequencing-based algorithms for detection of structural variations and discuss the key issues in future development.

Keywords: *copy number variations; paired-end sequencing; chromosomal alterations; translocations; indels*

INTRODUCTION

The human genome contains a diverse array of genomic variants. Among the most well-known are single nucleotide polymorphisms (SNPs), length polymorphisms of microsatellite sequences, and several types of structural variations (SVs). SVs include dosage-altering variations such as insertions and deletions, and dosage-invariant rearrangements such as inversions and translocations. Deletions and insertions larger than 1 kb [1] are often collectively referred to as copy number variations (CNVs), while smaller (<1 kb) insertions or deletions are referred to as indels. SNPs have long been thought to be the most common class of genetic variations and have been used widely in linkage and genome-wide association studies [2]. However, it is now recognized that other types of variations are also widespread in human genomes [3], even in the genomes of phenotypically normal individuals [4]. The database of genomic variants (DGV), for instance, lists about 60 000 CNVs, 850 inversions and 30 000 indels identified in healthy individuals (http://projects.tcag.ca/variation/; 25th March 2010 update).

The impact of SVs has been demonstrated in a wide range of applications including disease association studies, cancer genomics, and evolutionary studies [5–8]. Copy number changes, especially those involving genes sensitive to a dosage effect, are likely candidates that may result in phenotypic consequences. For example, initial SV studies successfully identified common CNVs in coding regions associated with several complex disease phenotypes such as autoimmune and infectious disorders [9, 10] as well as those associated with behavioral variation [11]. Small-scale deletion or duplication of conserved regulatory regions can affect the function of *cis*-regulated genes leading to a disease phenotype, as shown in the case of *DAX1* [12] and *SOX9* [13]. In recent large-scale disease association studies, rare but statistically

Corresponding author. Peter J. Park, Center for Biomedical Informatics, Harvard Medical School, 10 Shattuck St, 4th Floor, Boston, MA 02115, USA. Tel: 617-432-7373; Fax: 617-432-0693; E-mail: peter_park@harvard.edu

Ruibin Xi is a postdoctoral research associate at the Center for Biomedical Informatics, Harvard Medical School. He is interested in development of computational tools for analysis of next-generation sequencing data and applications of statistics and bioinformatics methods to cancer genetics and epigenetics.

Tae-Min Kim is a postdoctoral research associate at the Center for Biomedical Informatics, Harvard Medical School. He is interested in developing bioinformatics methods for cancer and population genetics.

Peter Park is an Assistant Professor at Children's Hospital Boston and Harvard Medical School. His laboratory focuses on computational analysis of high-throughput genomic data and its applications to cancer genomics and epigenomics.

significant SVs were identified for complex diseases such as autism and early onset obesity [14, 15]. With the identified SVs encompassing or near the disease-susceptibility genes, these studies not only provide potential disease markers but also elucidate the genetic architecture of genomic disorders and complex disease traits.

Until recently, microarray-based platforms were widely used to identify CNVs. Two pioneering studies used bacterial artificial chromosome (BAC) and oligonucleotide-based microarray comparative genomic hybridization (array-CGH) [16, 17]; the first generation genome-wide human CNV map was also constructed using these platforms [18]. However, BAC-based approaches cannot detect small CNVs or accurately map the boundaries of CNVs due to the large size of BACs [19]. Even for the newer oligonucleotide-based arrays containing more than 1 million probes, the resolution is still limited to 10–20 kb [20, 21]. Array-CGH also has several technical limitations, including intrinsic noise due to cross-hybridization and a limited dynamic range, and cannot detect dosage-invariant changes such as chromosomal translocations or inversions. As an alternative to hybridization-based methods, Sanger sequencing was also applied to identify genomic variants in normal individual genomes, for example using fosmid libraries [22, 23]. But the low throughput and high cost of Sanger sequencing imposed severe limitations on the number and size of detected SVs. For example, Kidd *et al.* [23] sequenced about 1 billion base pairs per individual for genome-wide SV discovery and identified about 4000 SVs for eight individuals. This level of sequencing by the Sanger method was too expensive and time-consuming for general use.

Next-generation sequencing (NGS) has enabled cost-effective, high-throughput sequencing [24]. The NGS platforms, first Roche 454 and later Illumina/Solexa Genome Analyzer and Applied Biosystems (ABI) SOLiD, generate orders of magnitude more sequences than the standard gel capillary-based technology. For instance, HiSeq2000, the latest model from Illumina, allows the researchers to obtain 30× coverage data for two human genomes in a single run. The NGS technology has been employed in all major areas of genetics and genomics. Among the major consoritum projects enabled by this technology are the 1000 genome project (http://www.1000genomes.org/), which aims to provide a comprehensive catalog of human genetic variation by sequencing a large number of people, and The Cancer Genome Atlas (TCGA) project (http://cancergenome.nih.gov/), which aims to generate a multi-dimensional genomic characterization of major tumor types.

These NGS platforms have also been used to examine SV. In a pioneering SV study using NGS, Korbel *et al.* [25] sequenced over 5 billion base pairs from two human genomes using the Roche 454 platform and identified 892 indels, 122 inverstions and 283 translocations. Since then, many have investigated SVs using NGS data with various algorithms. Most current SV detection algorithms adopt the 'comparison-versus-reference' strategy, in which they first align the short sequencing reads from the genome of interest to a known reference genome and then analyze the mapping signatures that could indicate SVs.

One simple feature to consider is the tag density along the genomic coordinates. Regions with more reads than expected would indicate copy gains in the sequenced genome, and vice versa for copy losses. The signature left by dosage-invariant SVs are more complex and generally cannot be detected by single-end sequencing used for tag counting. In the past year or two, however, the paired-end sequencing technology and protocols have become mature enough to be commonly used for detection of SVs. Briefly, paired-end reads (called 'mate pairs' when there is a long insertion in between) are generated by sequencing from both ends of the DNA library fragments whose sizes are approximately known (insert size). Some paired-end sequencing protocols involve circularization of the DNA segments that can generate paired-end reads with a larger insert size (usually several kilobase). The other way is to directly sequence both ends of the size-selected DNA fragments, generating paired-end reads with tighter insert size. The advantage of a large insert size in the first technique is that it is better at detecting large SVs. The second technique, on the other hand, provides higher resolution and is more powerful for detecting smaller SVs.

In this article, we will review currently available SV-detecting algorithms that utilize NGS data (Table 1). These algorithms can be classified into two types according to the read-mapping signatures they use: algorithms that search for regions with abnormal tag counts and algorithms that survey the configurations of the paired-end mappings (PEMs). In the following, we will discuss these two classes of

Table I: SV detection algorithms

Algorithm	Data type	Control	Insertion	Deletion	Inversion	Inter-chromosomal translocation	Intra-chromosomal translocation	Copy change	Reference
EWT	Single end	No	✓ (>100 bp)	✓ (>100 bp)				✓	32
MrFast	Single end	No	✓	✓				✓	30
CNV-seq	Single end	Yes	✓	✓				✓	35
Seg-seq	Single end	Yes	✓	✓				✓	33
rSW-seq	Single end	Yes	✓	✓				✓	34
VariationHunter	Paired end	No	✓	✓	✓				37, 38
MoDIL	Paired end	No	✓ (10–50 bp)	✓ (10–50 bp)					39
BreakDancer	Paired end	No	✓	✓	✓	✓	✓		41
PEMer	Paired end	No	✓	✓	✓	✓	✓		42
GASV	Paired end	No	✓	✓	✓	✓	✓		43
HYDRA	Paired end	No	✓	✓	✓	✓	✓		40
Pindel	Paired end	No	✓ (1–20 bp)	✓ (1–10 bp)					50
NovelSeq	Paired end	No	✓ (novel insertion)						36

algorithms in two separate sections. Then, we will address the future research directions and conclude the article.

ALGORIGHMS BASED ON TAG DENSITY

Most of the tag density-based algorithms assume that the sequence reads are single-end reads, though they may also be applied to paired-end data. The general strategy of these algorithms is to search for genomic regions whose tag counts are significantly different from the expected counts. In an ideal case, if a 'control' genome is available, the expected number is estimated from that genome. If there is no such genome, an assumption has to be made on the expected distribution of reads (see below). These tag density-based algorithms perform best at detecting large CNVs. As the CNV size increases, both sensitivity and specificity of the algorithms increase. However, these algorithms generally can only detect dosage-altering SVs such as indels and CNVs and cannot detect dosage-invariant SVs such as inversions and translocations.

Methods using depth of coverage

In a method based on depth of coverage (DOC), a genome is usually partitioned into non-overlapping windows and the tag counts in the windows are taken as a measure of DOC. Candidate SVs are usually determined as consecutive genomic windows in which the observed DOC is substantially different from the expected. The basic assumption of DOC-based methods is that the sequencing reads are randomly sampled with equal probability from any location on the sequenced genome. Thus, after aligning these reads to the reference genome, the read density of a given genomic window should be proportional to the local copy number. DOC-based methods are more cost-effective compared with the methods that need a control genome. However, significantly high or low read density is not necessarily caused by copy number changes since many other factors impact the local tag density. For example, current NGS platforms have GC-bias [26], some genomic regions are also more open and thus are fragmented more easily, and mappability of short reads is not constant across the genome [27]. These factors can affect local sequencing depth in the absence of actual copy number differences. The effect due to variation in mappability can be corrected computationally but other effects are more difficult to mitigate.

Many investigators have used DOC-based methods for SV discoveries [28–32]. For example, Yoon *et al.* [32] developed an algorithm called event-wise testing (EWT) for SV identification based on DOC. Their algorithm first counts tags in 100bp windows, corrects the tag counts for GC content and uses the adjusted tag count as the measure of DOC. Then, the authors convert the adjusted tag counts to Z-statistics and assign an upper-tail (for duplication) and a lower-tail (for deletion) P-value to each window. An interval consisting of l windows is called an unusual event for duplication (deletion) if all upper-tail (lower-tail) P-values of its l windows

are less than a specified threshold. This threshold is constant for the intervals of the same size, but it increases as the interval gets larger (increase in *l*). Lastly, EWT reports these events after some post-processing. The authors claimed that, for their 30× coverage data, the use of 100bp window allowed them to approximate the distribution of tag count by normal distribution without sacrificing too much resolution. But the choice of window size for data sets with different coverage was not discussed. Without a control genome, however, this approach has limited power to detect small insertions and deletions.

Algorithms using case versus control data set

The SV detection algorithms that utilize the reads from a control genome suffer less from the sequencing biases such as those introduced by variable GC content of genomic fragments, since one could reasonably assume that these biases for case and control genome are similar and that they should be cancelled out when the case is compared with the control. The general strategy of this class of algorithms is to look for the genomic regions in which there are significantly more (or fewer) case reads than the control reads (Figure 1). This strategy can be used for identification of disease-related genomic variants (e.g. comparison of tumor genome versus matched normal genome) [33, 34] as well as for comparison between normal genomes for SV screening [35]. The disadvantage of sequencing a control genome of course is that it doubles the cost.

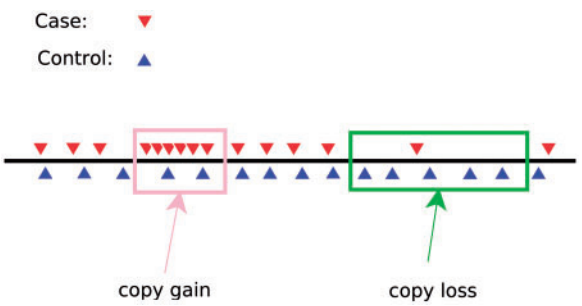

Figure 1: CNV detection using single-end reads with a control (reference) genome. The triangles represent mapped read positions along the case and control genomes. Regions in which the tag counts are different with statistical significance are identified as potential CNVs.

One common method for analyzing case/control sequencing data is to partition the genome into small windows of fixed size and use multiple testing procedures to find the windows in which the case genome has significantly greater or fewer reads than the control. The choice of the window size is essential for this method. Too large a window would sacrifice the resolution; too small a window would not give enough power for detecting low copy change regions. Xie and Tammi [35] proposed an algorithm CNV-seq to calculate the best window size given a significance level, a $\log_2$ copy ratio threshold and the coverage of the data. Their algorithm models the number of short reads in a genomic region as following a Poisson distribution. When the best window size is determined, CNV-seq identifies the windows in which copy ratios between the case and control are significantly different from 1. Since CNV-seq linearly scans the genome using sliding windows of fixed size, it is computationally fast. The estimation of parameters in the Poisson model used by CNV-seq essentially assumes a uniform distribution of the tags on the reference genome and that the only factor that affects the distribution of the tags is the copy number. Therefore, a Poisson distribution alone is too simple a model for the distribution of tags and more sophisticated statistical models are required.

The algorithms based on a multiple testing procedure such as CNV-seq usually only identify windows with significant differences in tag counts. One has to use a further window-merging step to estimate the breakpoint positions and distinguish focal CNVs from the surrounding large CNVs. Having fixed-size windows would miss some small but significant copy change regions. Chiang *et al.* [33] instead adopted another strategy and proposed an algorithm called SegSeq**.** They focused on the identification of breakpoints rather than CNV regions. Similar to CNV-seq, SegSeq also assumes the numbers of tumor (case) and normal (control) reads in a given genomic region follow Poisson distributions. Given a position that has at least one tumor read, SegSeq compares its left and right neighboring windows and calculates a *P*-value of copy-ratio change based on the Poisson model. If the *P*-value is less than a pre-specified significance level P_{ini}, the position will be viewed as a potential breakpoint. Notably, the neighboring windows are chosen to contain a fixed number of normal reads (default is 400 normal reads). As a result, the window size will be smaller

(larger) in the genomic regions with more (less) normal reads. Then, SegSeq iteratively compares the segments demarcated by the candidate breakpoints and joins the segments that are not significantly different at a new significant level P_{merg} The *P*-value in the merging step is calculated based on the total number of tumor and normal reads in the segments rather than the reads in the local windows, and hence it is generally different from the *P*-value calculated in the local change point identification step. The advantages of SegSeq include its ability of detecting focal CNV events embedded in a larger CNV event and its precise estimation of breakpoints. But, the *P*-value is also computed based on Poisson model and hence factors other than copy number change can result in a highly significant *P*-value. SegSeq also has three parameters that require tuning and the results can be sensitive to the parameters.

ALGORITHMS FOR SV DETECTION USING PEM

The general strategy for SV detection using PEMs is to align the paired-end reads from a sequenced genome onto the reference genome and look for 'discordant' PEMs that may indicate the presence of SVs nearby. A concordant or discordant mapping of the two ends of a paired-end read is determined by their mapped orientations and by comparing their mapped distances with the known insert size. This strategy was first carried out on a large-scale by Tuzun *et al.* [22] and has been adopted by most SV detection algorithms using PEMs. Advantages of PEM-based methods include the ability to detect dosage-invariant SVs, higher sensitivity for detecting smaller SVs, and the precision of localizing the breakpoint. But these algorithms have limited power in detecting insertions larger than the insert size. In the following, we will first discuss the configurations of PEMs at the presence of various types of SVs and then review the current algorithms based on PEMs.

Configurations of discordant PEMs

Different SVs can lead to different classes of discordant paired-end reads. The most obvious discordant PEMs are those with their two ends mapped to different reference chromosomes (Figure 2F), most likely reflecting an inter-chromosomal translocation. Other types of discordant PEMs are less obvious and require a comparison of the distance between the two mapped positions of a read to its expected distance based on the insert size. The target insert size is known in a given experiment, but the actual sizes are variable due to experimental noise. The distribution of the distances nonetheless can be estimated from the mapped distances on the reference genome [22]. If the distance between the two ends on the reference is significantly larger than the estimated insert size, the read is likely to contain a deletion (Figure 2A). Similarly, the pair whose mapped distance is significantly smaller than the insert size is likely to reflect an insertion on the sequenced genome (Figure 2B). A paired-end read that spans one of the breakpoints of an inversion could be distinguished by an 'incorrect' orientation of the end lying in the inversion (Figure 2C).

Tandem duplication can also lead to discordant paired-end reads. Suppose that a paired-end read spans the breakpoint of a tandem-duplicated region (Figure 2D) and the insert size is relatively small. Then, we would observe a paired-end read whose orientations are mapped correctly but the order of the two ends is reversed. Figure 2E shows two types of discordant paired-end reads that can result from an intra-chromosomal translocation. One type of paired-end read (in deep blue) behaves as a discordant paired-end read coming from a deletion, and the other type (in light blue) as a read from a tandem duplication. In some cases, the configurations of the PEMs can be misleading. For example, if there is a tandem duplication and the insert size of the paired-end read is larger than the duplicated sequence, the paired-end reads that span the breakpoints would have mapped distances significantly less than the insert size (Figure 3A and B) and hence the configuration of the PEM would be similar to the discordant PEMs from an insertion. But this could be misleading since one may mistakenly conclude that there is an insertion between the pair (Figure 3A).

In practice, PEMs may have even more complicated configurations than discussed above. For example, an insertion in the sequenced genome may not exist in the reference genome, and this can generate a paired-end read whose one end can be mapped to the reference genome but the other end cannot be mapped (Figure 2G). Furthermore, combinations of the six basic structural variants discussed above could exist in the sequenced genome. These more complex configurations of

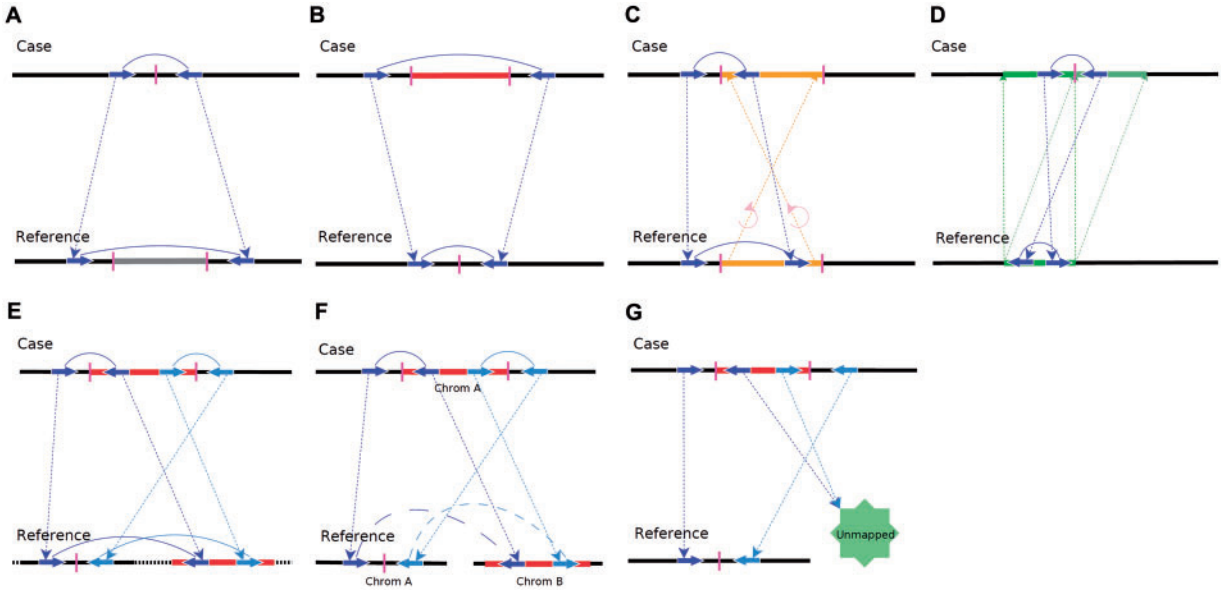

Figure 2: Configurations of PEMs in various types of SVs. (**A**) Deletion. The paired-end read spans the breakpoint of a deletion. Thus, the mapped distance of the paired-end reads is significantly larger than the insert size. (**B**) Insertion. The paired-end reads spans an insertion, and the mapped distance significantly less than the insert size. (**C**) Inversion. The read pair encompasses one breakpoint of an inversion and the right end is mapped with incorrect orientation. (**D**) Tandem duplication. The read pair spans the middle breakpoint of a tandem duplication. The PEM will have correct orientation but with reverse order. (**E**) Intra-chromosomal translocation. Two read pairs span the two breakpoints of an intra-chromosomal translocation with one pair having a large mapped distance and the other having correct orientation but their ordering reversed. (**F**) Inter-chromosomal translocation. The two ends of the pair are mapped to different chromosomes. (**G**) One-end unmapped. The sequenced genome has a DNA segment that does not exist in the reference genome. One end of the pair is mappable but the other is not.

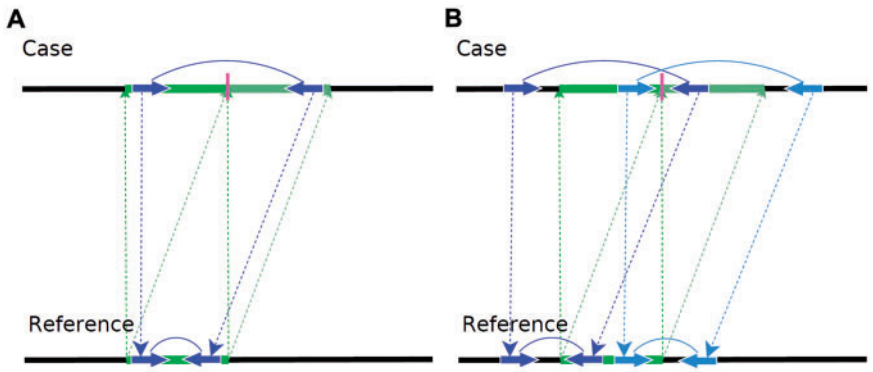

Figure 3: The insert size is larger than the duplicated segment in a tandem duplication. (**A**) The read pair spans the middle breakpoint. This could be misleading because one may mistakenly conclude that there is an insertion between the pair. (**B**) The read pair spans one of the duplicated segments.

discordant PEMs would be much more difficult to disentangle.

Algorithms for analyzing the PEMs

Many PEM-based algorithms for SV detection have been proposed [36–43]. These algorithms can be put into two categories: clustering-based algorithms and distribution-based algorithms. The idea behind clustering-based algorithms is to classify the PEMs as concordant and discordant pairs and then to cluster the discordant PEMs based on the positions of the paired-end reads. The discordant PEMs are the pairs with incorrect orientations or the pairs with mapped distances beyond a fixed range. For example, Tuzun *et al.* [22] classified a PEM as a discordant pair if its mapped distance is three SDs away from the mean

insert size, while Korbel *et al.* [42] used simulations to determine the expected range of concordant PEMs. The discordant PEMs are clustered together if they have the same type of configuration and span the same SV. Then, those clusters with more than a specified number of discordant PEMs (usually two or more) are taken as potential SVs.

The early clustering-based methods only considered uniquely mapped paired-end reads and removed the reads that have multiple alignments before clustering. Hence, these methods generally cannot detect SVs in the genomic regions rich in repeats and segmental duplications. Recently, Hormozdiari *et al.* [37] proposed an algorithm called VaraitionHunter that utilizes the multiple alignable reads and hence allows detection of SVs in the repeat and segemental duplication-rich regions. VariationHunter first uses MrFast [30] to get all possible alignments of each read. A paired-end read will be classified as discordant if none of its multiple mappings is concordant. Then, VariationHunter clusters the alignments of the discordant paired-end reads. At this step, different alignments of one paired-end read can belong to different clusters and therefore support different SVs. To solve this problem, Hormozdiari *et al.* proposed to assign a 'best' alignment to the discordant paired-end read via minimizing the total number of implied SVs, or, in other words, via solving the maximum parsimony structural variation (MPSV) problem.

Since the clustering-based methods classify the mappings as discordant PEMs based on a fixed range of mapped distances, they can only detect relatively large indels. For example, if the fixed range is chosen as within 3 SDs from the mean insert size, the insertion of size about 2 SDs will not be identified, since the paired-end reads encompassing this insertion would have mapped distances about 2 SDs away from the mean insert size and their mappings will be classified as concordant. To address this question, Lee *et al.* [39] proposed to detect these smaller SVs based on the distribution of the mapped distances. The idea of their algorithm MoDIL is to compare the local distribution of the mapped distances to the genome wide distribution of insert sizes [called $p(y)$]. Given a genomic locus i, if there were no indel nearby, the local distribution of the mapped distance [called $p(C_i)$] would be identical to the distribution $p(y)$. However, if there were a homozygous insertion or deletion near the locus i, the distribution $p(C_i)$ would be a shift of the distribution $p(y)$ (Figure 4A). In the case of a heterozygous indel, the distribution $p(C_i)$ will be a mixture of the shifts of the distribution $p(y)$ (with different magnitude or direction) (Figure 4B).

The clustering-based methods and distribution-based methods have little overlap. The clustering-based methods are good at detecting large translocations, inversions and relatively larger indels; the distribution-based methods are better at detecting smaller indels and cannot detect translocations and inversions. Chen *et al.* [41] proposed to combine the two methods and showed that they can detect a wider range of SVs. Their algorithm BreakDancer consists of two complementary algorithms BreakDancerMax and BreakDancerMini. BreakDancerMax is a clustering-based algorithm

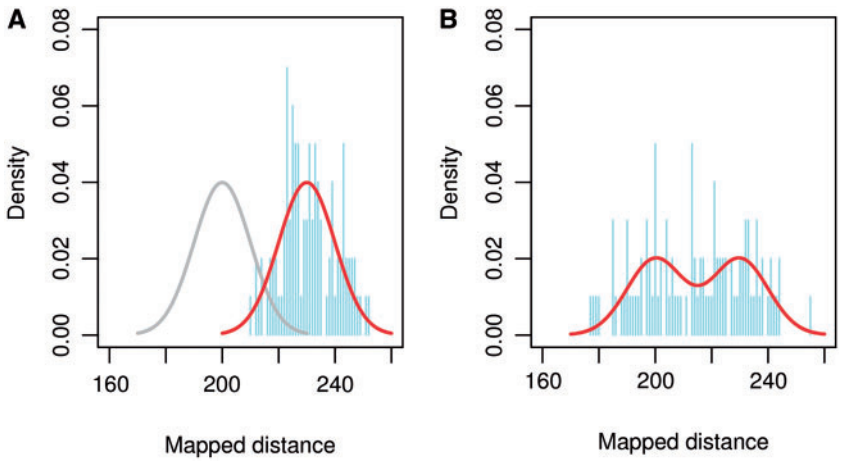

Figure 4: Local distribution of mapped distances of the paired-end reads. (**A**) A homozygous deletion. The distribution is shifted to right, where the gray curve is the global distribution of the mapped distance. (**B**) A heterozygous deletion. The distribution is a mixture of two distributions and therefore is a bimodal distribution.

and is designed for detecting larger translocations, inversions and larger indels, and BreakDancerMini is designed to detect small indels that cannot be identified by BreakDancerMax. Given that no single algorithm is likely to detect all types of SVs equally well, a judicious combination of multiple complimentary methods is likely to be a fruitful approach.

DISCUSSION

The performance of these algorithms depends on the accuracy and completeness of the aligned reads in representing the underlying structure of SVs. There are at least four major issues associated with the current NGS data sets. The first is the elevated sequencing error rate of NGS platforms, which is still higher than the conventional Sanger sequencing, especially at the first one or two positions and increasing exponentially near the end of the read. The sequencing errors cause a substantial loss of reads during alignment [26]. Second, given the short length of the sequenced bases, many reads do not map uniquely to the genome. In particular, nearly half of the human genome consists of repetitive sequences but the reads from the repeat-rich regions that align to multiple locations are typically ignored. Thus, SVs in these regions are often not effectively represented by uniquely alignable reads. Instead of ignoring the reads with multiple alignments, it is also possible to build various models to assign a 'best' position to these reads and use them for SV discovery, as done, for example, in VariationHunter and MoDIL. But the effect of these strategies has not been carefully evaluated so far. The short length of the reads also results in uneven mappability of regions along the genome as described earlier. Third, as also described in the context of tag counting methods, certain regions of the genomes are represented at a higher rate than others, while others are the opposite. This is due to the GC bias in sequencing steps [26], amplification errors (if used), and an uneven likelihood of fragmentation along the genome. Finally, many of the data sets do not have sufficient coverage to infer all SVs with statistical significance. The depth of sequencing necessary for reasonable values of sensitivity and specificity has not been examined so far. With low coverage data sets, the sensitivity of the SV detection is limited, and algorithms for SV detction in this context such as the one developed by Lee *et al.* [44] will be important for studies with large cohorts.

There are also a number of situations in which it is difficult to infer the correct genomic configuration. For example, if the sequenced genome contains a large insertion (larger than the read length and insert size) that does not exist in the reference genome, the short reads sequenced from this region will be unalignable. In this case, *de novo* genome assembly might be required to obtain the individual's specific sequences that do not exist in the reference genome. A number of genome assembler using short reads have been proposed [45, 46]; however, it is still difficult to reconstruct a repeat-rich, diploid human genome using short reads from current NGS platforms. Recently, Hajirasouliha *et al.* [36] proposed an algorithm called NovelSeq to identify long novel insertions based on *de novo* assembly of unmapped reads. The unmapped reads are first classified into 'both-end unmapped' reads and only 'one-end unmapped' reads. The one-end unmapped reads are further clustered according to the mapped position and strand of the mapped ends. Then, *de novo* assembly is performed on the both-end unmapped reads and each cluster of one-end unmapped reads, making the computational time significantly reduced compared with performing *de novo* assembly on all the reads. At last, the assembled contigs are reported as potential novel sequence insertions in the genome.

Furthermore, if a read is sampled across the breakpoint of a deletion (Figure 5A) or across a small insertion (Figure 5B), that read also is not alignable to the reference. In this case, a split mapping could be employed, in which the read encompassing a breakpoint may be split correctly and mapped to

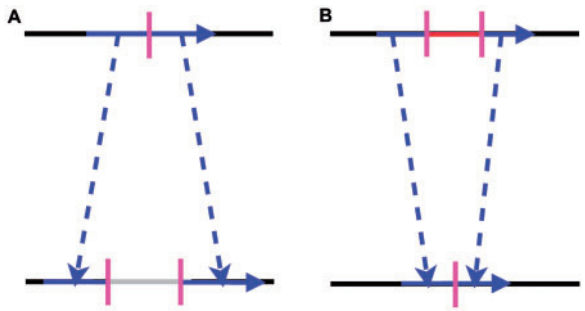

Figure 5: Split mapping. (**A**) The read that spans the breakpoint of a deletion can be split and mapped to two positions on the reference genome. (**B**) The read that contains a small insertion is mapped to the reference genome after removing the small insertion.

different positions (Figure 5). A recent study [47] successfully identified fusion genes using split mapping of single-end read data. Although the single-end-based approaches are clearly not as effective as the paired-end-based methods in general [48], methods initially developed for single-end reads can still be applied in some contexts, including the detection of splicing events in RNA-seq data [49]. Recently, Ye *et al.* [50] showed that if one end of the mate pair can be uniquely mapped to the reference genome but the other cannot be mapped, one can align the unmappable end onto the reference genome using split mapping and identify small indels. With one end already mapped and insert size known, the search space for split mapping for the other end becomes much smaller and the mapping can be done efficiently. It is of note that as the read length becomes longer, the PEM-based strategies will be applicable to single-end data, i.e., one can take the two ends of a single read, treat and analyze them as a paired-end read. In this case, the insert size is known exactly and the determination of the discordant pairs would be straightforward. Therefore, this method should give more precise prediction of deletions and insertions, and have a lower false positive rate.

For the PEM-based methods, the insert size is an important parameter for SV detection. With the same sequencing coverage, a larger insert size will give higher physical coverage and hence will be more cost-effective in detecting larger events. But, a smaller insert size will be more sensitive at detecting smaller events and in localizing breakpoints. Bashir *et al.* [51] studied the insert size effect for the gene fusion detection using PEMs, but more extensive investigations on this topic are necessary in other contexts. Other than single-end and paired-end sequencing, Ritz *et al.* [52] proposed an algorithm that uses strobe sequencing technology developed by Pacific Biosciences to detect SVs. This sequencing technology, though not widely available yet, can generate 'strobe reads' consisting of multiple subreads from a single DNA fragment. A strobe read with two subreads is similar to a paired-end read, but a strobe read with multiple subreads contains more information about the genome than a paired-end read. Therefore, strobe reads should allow one to detect SVs with higher sensitivity and lower false positives, as was shown in [52].

The tag density based methods and PEM-based methods are two complementary approaches. While the tag density based methods are better at detecting larger dosage-altering SVs, the PEM-based methods are better at smaller SV and dosage-invariant detection. Volik *et al.* [29] and Campbell *et al.* [31] used both approaches for SV detection, but they only applied them independently and did not combine them to achieve higher specificity or to increase the precision of breakpoints. As very long reads and paired-end sequencing are becoming technically feasible on many platforms, a method that combines the two complementary approaches can make more confident SV calls than the methods based only on one approach, e.g. verifying a putative deletion discovered by a PEM-based method by tag density in that region. Finally, it is important to note that there are many inefficiencies and a lack of rigor in these methods. Given the rapid developments in sequencing technology, these methods will have to be continually modified and improved in order to learn more about SVs and their consequences.

Key Points

- SVs constitute a substantial fraction of genetic variation in human genome, which may serve as valuable genetic markers in clinical and evolutionary studies.
- NGS provides opportunities for genome-wide SV assay with higher resolution and larger categories of variations than the conventional microarray-based methods.
- Current SV detection algorithms using NGS data can be categorized according to the read-mapping signatures they use: tag density-based algorithms and PEM-based algorithms The tag density-based methods such as EWT, CNV-seq and SegSeq measure copy number changes by the observed versus expected read counts (compared with a control genome or a model of underlying tag distribution) in windows along chromosomes and identify the regions with abnormal tag counts as potential CNVs. The PEM-based methods, including VariationHunter, MoDIL, BreakDancer and PEMer, make the SV calls by analyzing the configurations of the discordant PEMs.

FUNDING

This work was supported by the US National Institutes of Health (grants RC1 HG005482 and R01 GM082798 to P.J.P.).

References

1. Freeman JL, Perry GH, Feuk L, *et al.* Copy number variation: new insights in genome diversity. *Genome Res* 2006;**16**:949–61.
2. International HapMap Consortium. A haplotype map of the human genome. *Nature* 2005;**437**:1299–320.
3. Check E. Human genome: patchwork people. *Nature* 2005; **437**:1084–6.

4. Conrad DF, Pinto D, Redon R, *et al*. Origins and functional impact of copy number variation in the human genome. *Nature* 2010;**464**:704–12.
5. Buchanan JA, Scherer SW. Contemplating effects of genomic structural variation. *Genet Med* 2008;**10**:639–47.
6. Feuk L, Marshall CR, Wintle RF, *et al*. Structural variants: changing the landscape of chromosomes and design of disease studies. *Hum Mol Genet* 2006;**15**(Spec No 1):R57–66.
7. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* 2009;**10**:451–81.
8. McCarroll SA, Altshuler DM. Copy-number variation and association studies of human disease. *Nat Genet* 2007;**39**: S37–42.
9. Gonzalez E, Kulkarni H, Bolivar H, *et al*. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* 2005;**307**:1434–40.
10. Fanciulli M, Norsworthy PJ, Petretto E, *et al*. FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat Genet* 2007;**39**:721–3.
11. Perry GH, Dominy NJ, Claw KG, *et al*. Diet and the evolution of human amylase gene copy number variation. *Nat Genet* 2007;**39**:1256–60.
12. Smyk M, Berg JS, Pursley A, *et al*. Male-to-female sex reversal associated with an approximately 250 kb deletion upstream of NR0B1 (DAX1). *Hum Genet* 2007;**122**:63–70.
13. Kurth I, Klopocki E, Stricker S, *et al*. Duplications of noncoding elements 5' of SOX9 are associated with brachydactyly-anonychia. *Nat Genet* 2009;**41**:862–3.
14. Glessner JT, Wang K, Cai G, *et al*. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* 2009;**459**:569–73.
15. Bochukova EG, Huang N, Keogh J, *et al*. Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* 2010;**463**:666–70.
16. Iafrate AJ, Feuk L, Rivera MN, *et al*. Detection of large-scale variation in the human genome. *Nat Genet* 2004;**36**:949–51.
17. Sebat J, Lakshmi B, Troge J, *et al*. Large-scale copy number polymorphism in the human genome. *Science* 2004;**305**: 525–8.
18. Redon R, Ishikawa S, Fitch KR, *et al*. Global variation in copy number in the human genome. *Nature* 2006;**444**: 444–54.
19. Ylstra B, van dI, Carvalho B, *et al*. BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). *Nucleic Acids Res* 2006; **34**:445–50.
20. Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet* 2007;**39**:S16–21.
21. Cooper GM, Zerr T, Kidd JM, *et al*. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat Genet* 2008;**40**:1199–203.
22. Tuzun E, Sharp AJ, Bailey JA, *et al*. Fine-scale structural variation of the human genome. *Nat Genet* 2005;**37**:727–32.
23. Kidd JM, Cooper GM, Donahue WF, *et al*. Mapping and sequencing of structural variation from eight human genomes. *Nature* 2008;**453**:56–64.
24. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol* 2008;**26**:1135–45.
25. Korbel JO, Urban AE, Affourtit JP, *et al*. Paired-end mapping reveals extensive structural variation in the human genome. *Science* 2007;**318**:420–6.
26. Dohm JC, Lottaz C, Borodina T, *et al*. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* 2008;**36**:e105.
27. Rozowsky J, Euskirchen G, Auerbach RK, *et al*. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* 2009;**27**:66–75.
28. Bailey JA, Gu Z, Clark RA, *et al*. Recent segmental duplications in the human genome. *Science* 2002;**297**: 1003–7.
29. Volik S, Zhao S, Chin K, *et al*. End-sequence profiling: sequence-based analysis of aberrant genomes. *Proc Natl Acad Sci USA* 2003;**100**:7696–701.
30. Alkan C, Kidd JM, Marques-Bonet T, *et al*. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* 2009;**4**:1061–7.
31. Campbell PJ, Stephens PJ, Pleasance ED, *et al*. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* 2008;**40**:722–9.
32. Yoon S, Xuan Z, Makarov V, *et al*. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res* 2009;**19**:1586–92.
33. Chiang DY, Getz G, Jaffe DB, *et al*. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods* 2009;**6**:99–103.
34. Kim TM, Luquette LJ, Xi R, *et al*. rSW-seq: algorithm for detection of copy number alterations in deep sequencing data. *BMC Bioinformatics* 2010;**11**:432.
35. Xie C, Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* 2009;**10**:80.
36. Hajirasouliha I, Hormozdiari F, Alkan C, *et al*. Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics* 2010;**26**:1277–83.
37. Hormozdiari F, Alkan C, Eichler EE, *et al*. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res* 2009;**19**: 1270–8.
38. Hormozdiari F, Hajirasouliha I, Dao P, *et al*. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics* 2010;**26**:i350–7.
39. Lee S, Hormozdiari F, Alkan C, *et al*. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat Methods* 2009;**6**:473–4.
40. Quinlan AR, Clark RA, Sokolova S, *et al*. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res* 2010;**20**:623–35.
41. Chen K, Wallis JW, McLellan MD, *et al*. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* 2009;**6**:677–81.
42. Korbel JO, Abyzov A, Mu XJ, *et al*. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol* 2009;**10**:R23.

43. Sindi S, Helman E, Bashir A, *et al*. A geometric approach for classification and comparison of structural variants. *Bioinformatics* 2009;**25**:i222–30.

44. Lee S, Xing E, Brudno M. MoGUL: detecting common insertions and deletions in a population. *Res Comput Mol Biol* 2010;**6044**:357–68.

45. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 2008;**18**:821–9.

46. Butler J, MacCallum I, Kleber M, *et al*. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res* 2008;**18**:810–20.

47. Maher CA, Kumar-Sinha C, Cao X, *et al*. Transcriptome sequencing to detect gene fusions in cancer. *Nature* 2009; **458**:97–101.

48. Maher CA, Palanisamy N, Brenner JC, *et al*. Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc Natl Acad Sci USA* 2009;**106**:12353–8.

49. Ameur A, Wetterbom A, Feuk L, *et al*. Global and unbiased detection of splice junctions from RNA-seq data. *Genome Biol* 2010;**11**:R34.

50. Ye K, Schulz MH, Long Q, *et al*. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 2009;**25**:2865–71.

51. Bashir A, Volik S, Collins C, *et al*. Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS Comput Biol* 2008;**4**:e1000051.

52. Ritz A, Bashir A, Raphael BJ. Structural variation analysis with strobe reads. *Bioinformatics* 2010;**26**:1291–8.